视觉Transformer 年度进展评述

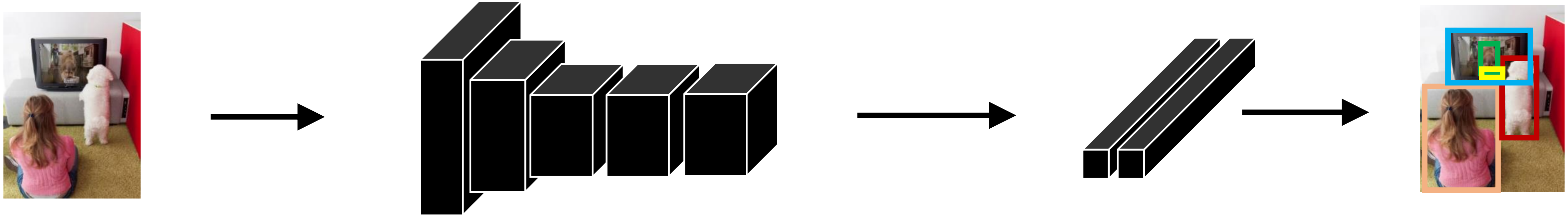
胡瀚
微软亚洲研究院

@VALSE 2022 天津
2022.08.22

2021——视觉Transformer年：**三大变革**

视觉骨干网络 & 预训练

头部网络



CNN时代

CNN
e.g., ResNet

有监督预训练
(图像分类)

针对不同任务的不同架构
e.g., Mask R-CNN, FCN

Transformer
时代

Transformer
e.g., ViT/Swin

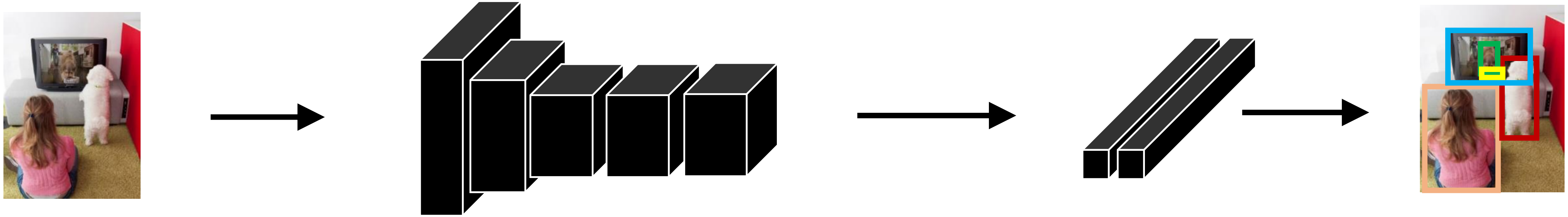
自监督预训练MIM：
e.g., BEiT/MAE/SimMIM

Transformer
e.g., DETR

2021——视觉Transformer年：视觉Transformer取代CNN（变革#1）

视觉骨干网络 & 预训练

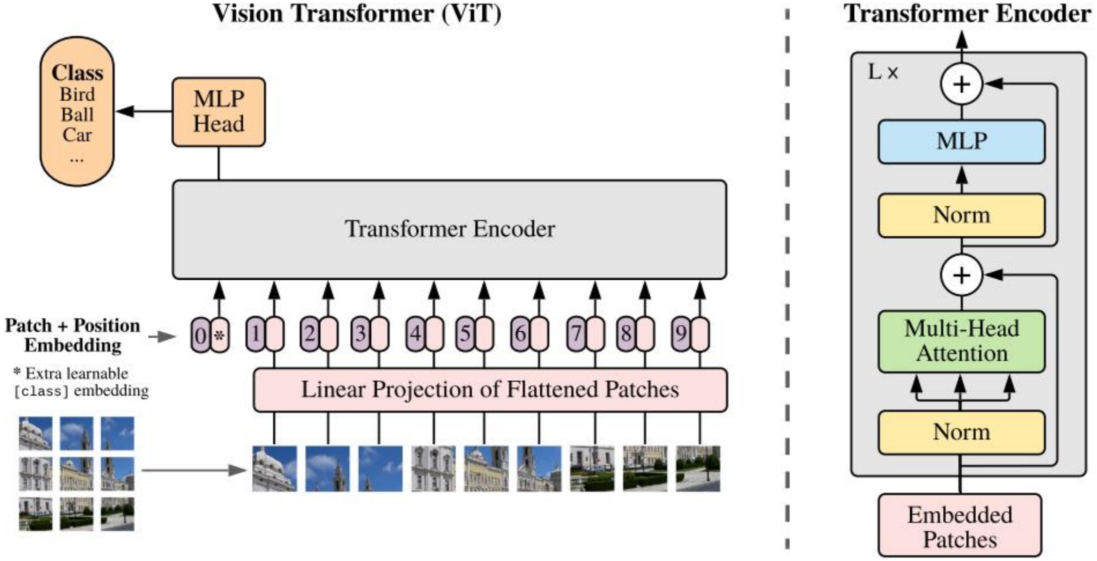
头部网络



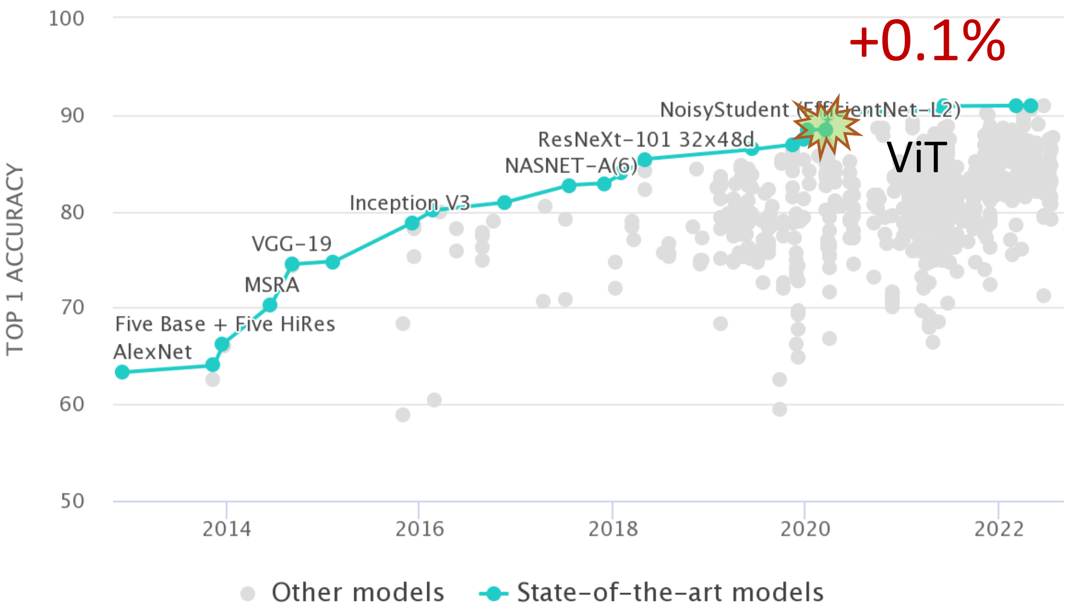
CNN时代	CNN e.g., ResNet	有监督预训练 (图像分类)	针对不同任务的不同架构 e.g., Mask R-CNN, FCN
Transformer 时代	**Transformer** **e.g., ViT/Swin**	**自监督预训练MIM：** **e.g., BEiT/MAE/SimMIM**	**Transformer** **e.g., DETR**

视觉Transformer骨干模型兴起

- **里程碑 (2020-10)**：视觉Transformer (ViT) [谷歌]
    - 图像分类取得突破
    - 相比CNN的核心优势：**动态性，长程建模能力**



ImageNet图像分类

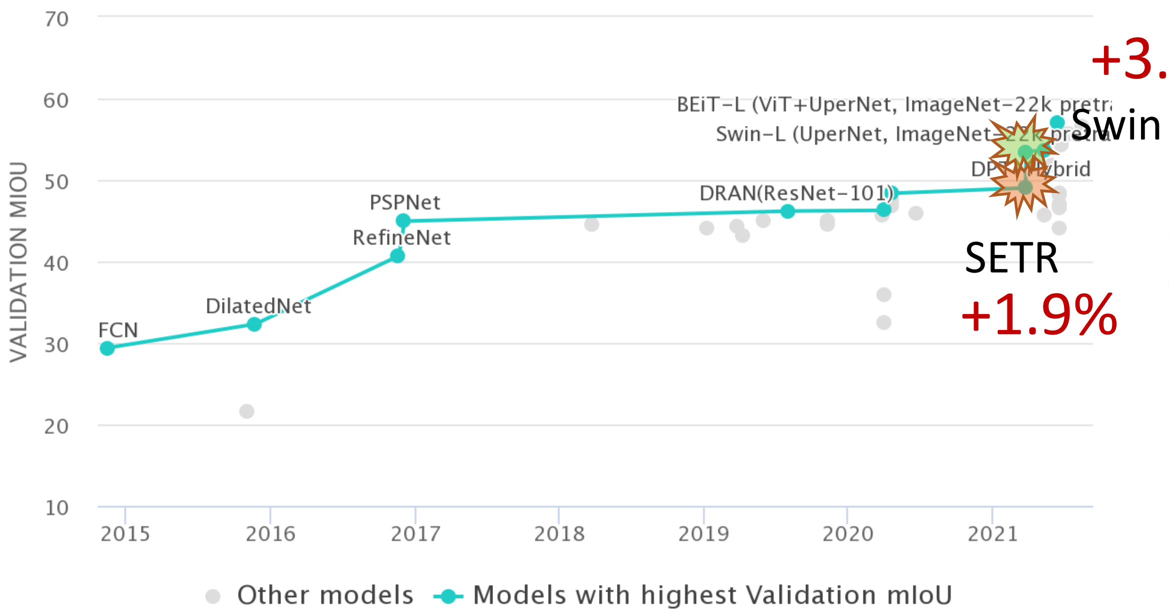


Alexey Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

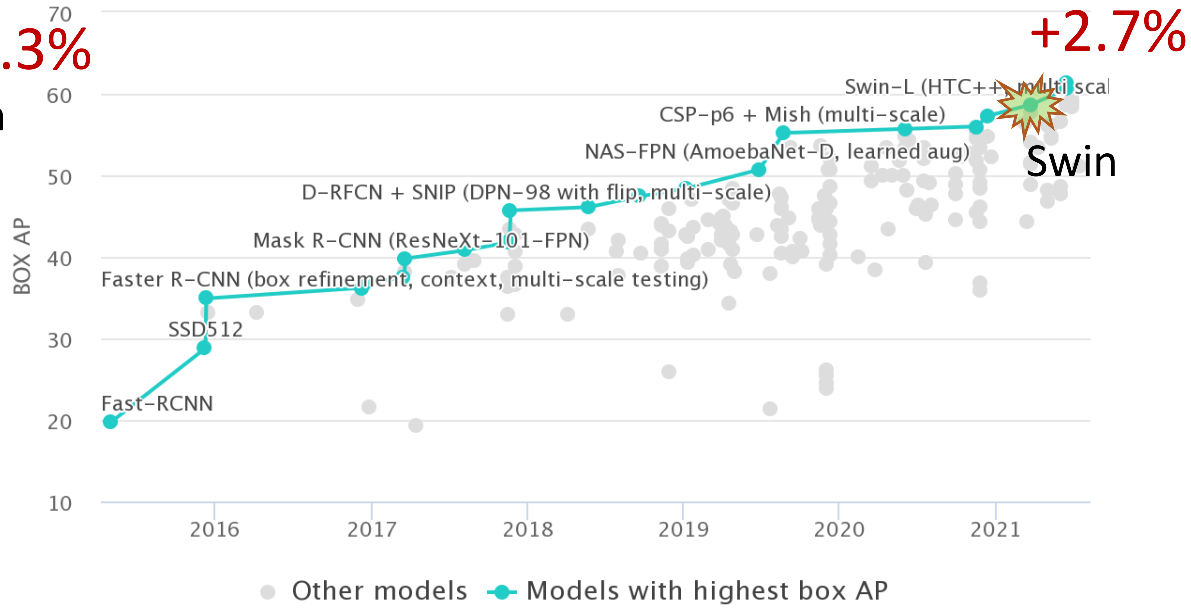
视觉Transformer骨干模型兴起

- 密集视觉任务取得新记录
  - SETR-ViT（分割） [复旦&牛津 CVPR2021]
  - Swin Transformer (检测、分割) [中科大&西交&微软 ICCV2021]

ADE20K语义分割



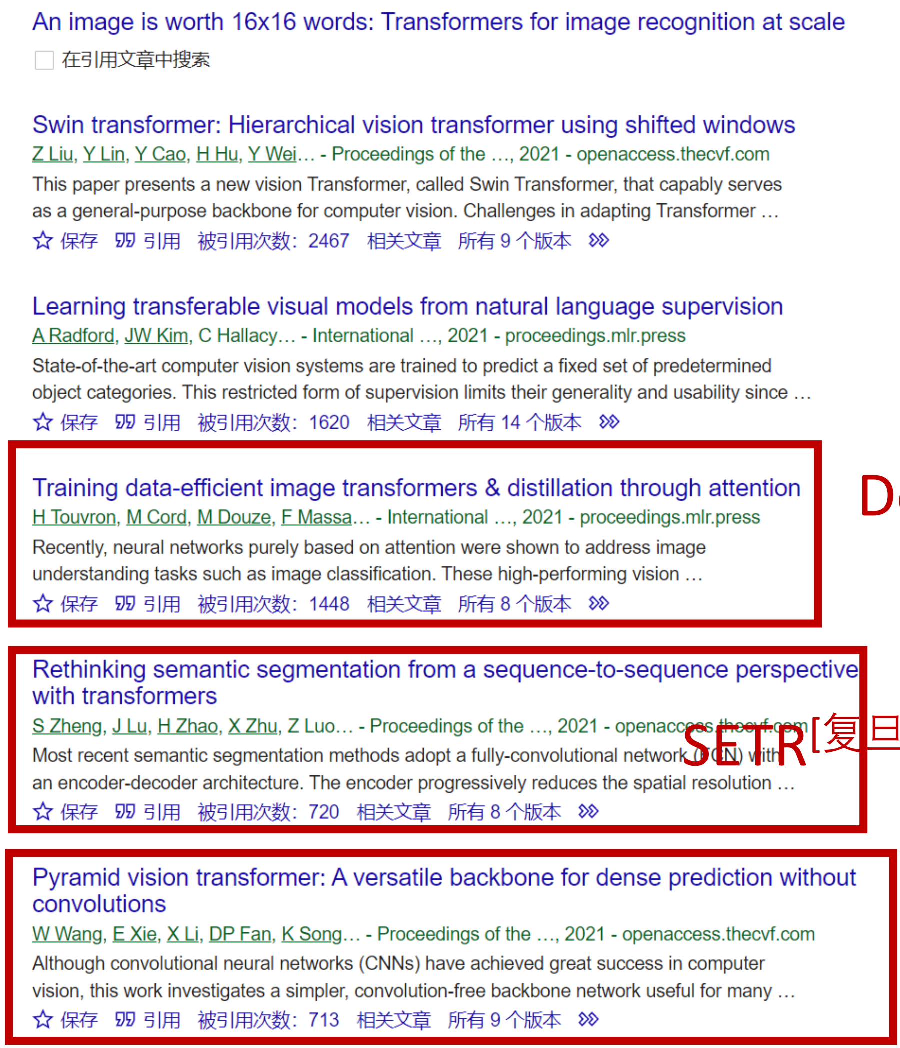
COCO物体检测



Sixiao Zheng et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers, CVPR 2021
Ze Liu et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, ICCV2021

大量ViT的Follow-Up工作

- 华人是主要贡献力量



An image is worth 16x16 words: Transformers for image recognition at scale

☐ 在引用文章中搜索

Swin transformer: Hierarchical vision transformer using shifted windows
Z Liu, Y Lin, Y Cao, H Hu, Y Wei… - Proceedings of the …, 2021 - openaccess.thecvf.com
This paper presents a new vision Transformer, called Swin Transformer, that capably serves as a general-purpose backbone for computer vision. Challenges in adapting Transformer …
☆ 保存 99 引用 被引用次数：2467 相关文章 所有 9 个版本 »

Learning transferable visual models from natural language supervision
A Radford, JW Kim, C Hallacy… - International …, 2021 - proceedings.mlr.press
State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since …
☆ 保存 99 引用 被引用次数：1620 相关文章 所有 14 个版本 »

Training data-efficient image transformers & distillation through attention
H Touvron, M Cord, M Douze, F Massa… - International …, 2021 - proceedings.mlr.press
Recently, neural networks purely based on attention were shown to address image understanding tasks such as image classification. These high-performing vision …
☆ 保存 99 引用 被引用次数：1448 相关文章 所有 8 个版本 »

DeiT[Meta]

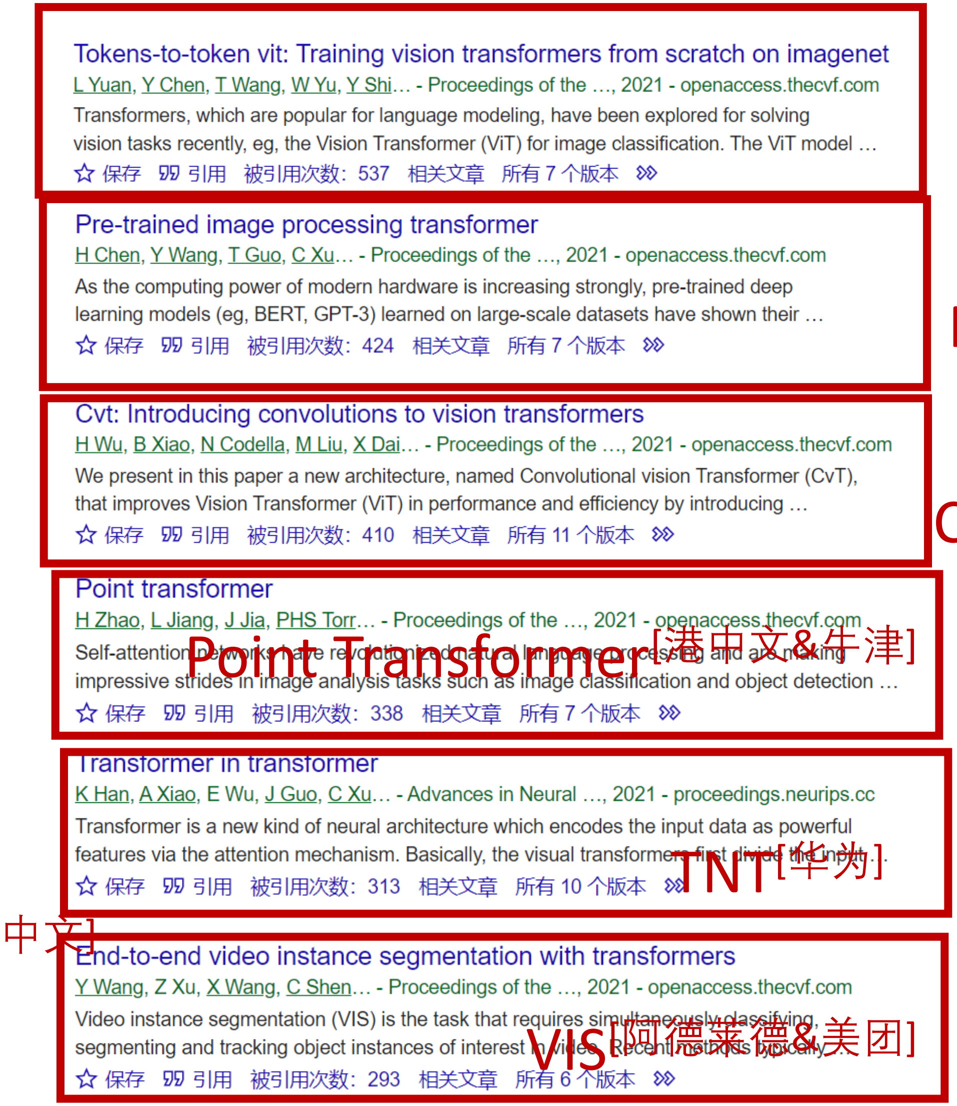
Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers
S Zheng, J Lu, H Zhao, X Zhu, Z Luo… - Proceedings of the …, 2021 - openaccess.thecvf.com
Most recent semantic segmentation methods adopt a fully-convolutional network (FCN) with an encoder-decoder architecture. The encoder progressively reduces the spatial resolution …
☆ 保存 99 引用 被引用次数：720 相关文章 所有 8 个版本 »

SETR[复旦&牛津&腾讯]

Pyramid vision transformer: A versatile backbone for dense prediction without convolutions
W Wang, E Xie, X Li, DP Fan, K Song… - Proceedings of the …, 2021 - openaccess.thecvf.com
Although convolutional neural networks (CNNs) have achieved great success in computer vision, this work investigates a simpler, convolution-free backbone network useful for many …
☆ 保存 99 引用 被引用次数：713 相关文章 所有 9 个版本 »

PVT[南大&港中文]

T2T-ViT[NUS]



Tokens-to-token vit: Training vision transformers from scratch on imagenet
L Yuan, Y Chen, T Wang, W Yu, Y Shi… - Proceedings of the …, 2021 - openaccess.thecvf.com
Transformers, which are popular for language modeling, have been explored for solving vision tasks recently, eg, the Vision Transformer (ViT) for image classification. The ViT model …
☆ 保存 99 引用 被引用次数：537 相关文章 所有 7 个版本 »

Pre-trained image processing transformer
H Chen, Y Wang, T Guo, C Xu… - Proceedings of the …, 2021 - openaccess.thecvf.com
As the computing power of modern hardware is increasing strongly, pre-trained deep learning models (eg, BERT, GPT-3) learned on large-scale datasets have shown their …
☆ 保存 99 引用 被引用次数：424 相关文章 所有 7 个版本 »

PIT[华为]

Cvt: Introducing convolutions to vision transformers
H Wu, B Xiao, N Codella, M Liu, X Dai… - Proceedings of the …, 2021 - openaccess.thecvf.com
We present in this paper a new architecture, named Convolutional vision Transformer (CvT), that improves Vision Transformer (ViT) in performance and efficiency by introducing …
☆ 保存 99 引用 被引用次数：410 相关文章 所有 11 个版本 »

CvT[微软]

Point transformer
H Zhao, L Jiang, J Jia, PHS Torr… - Proceedings of the …, 2021 - openaccess.thecvf.com
Self-attention networks have revolutionized natural language processing and are making impressive strides in image analysis tasks such as image classification and object detection …
☆ 保存 99 引用 被引用次数：338 相关文章 所有 7 个版本 »

Point Transformer[港中文&牛津]

Transformer in transformer
K Han, A Xiao, E Wu, J Guo, C Xu… - Advances in Neural …, 2021 - proceedings.neurips.cc
Transformer is a new kind of neural architecture which encodes the input data as powerful features via the attention mechanism. Basically, the visual transformers first divide the input …
☆ 保存 99 引用 被引用次数：313 相关文章 所有 10 个版本 »

TNT[华为]

End-to-end video instance segmentation with transformers
Y Wang, Z Xu, X Wang, C Shen… - Proceedings of the …, 2021 - openaccess.thecvf.com
Video instance segmentation (VIS) is the task that requires simultaneously classifying, segmenting and tracking object instances of interest in video. Recent methods typically …
☆ 保存 99 引用 被引用次数：293 相关文章 所有 6 个版本 »

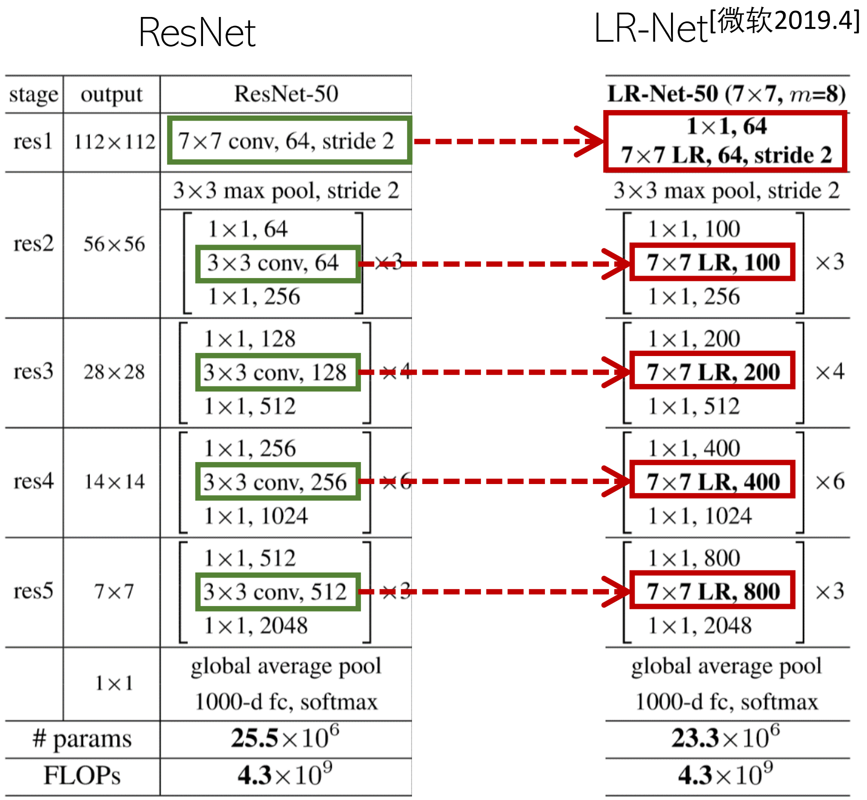
VIS[阿德莱德&美团]

更早期基于注意力和自注意力工作

- 全注意力视觉骨干网络
  - LR-Net [微软 2019.4]
  - SASA [谷歌 2019.5] （Transformer原作者团队）

- 与卷积骨干网络互补
  - Non-Local Networks [CMU&Meta 2017]
  - GCNet [微软 2019]/DANet [自动化所 2018] /OCNet [微软 2018]/CCNet [华科 2019]

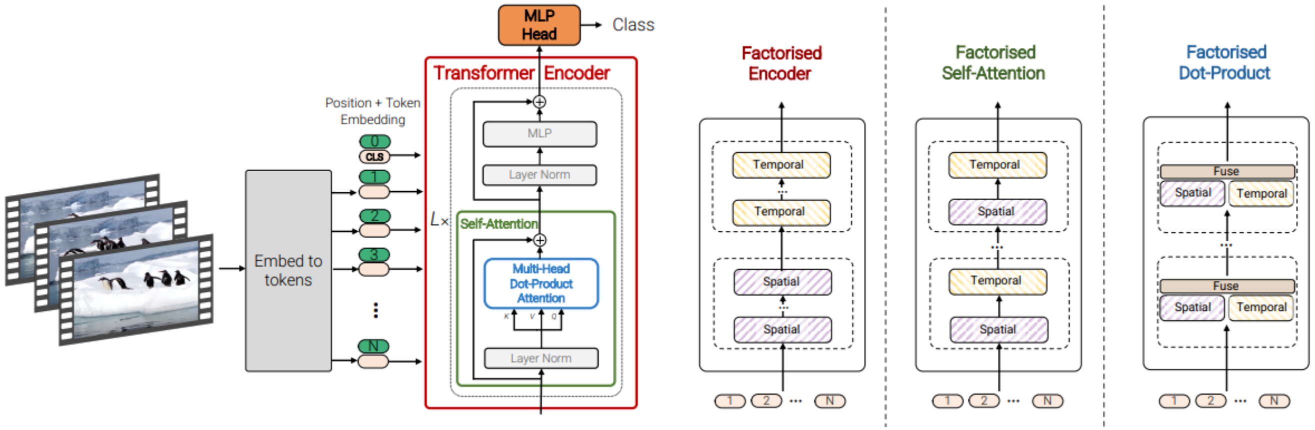
- 注意力应用于头部网络
  - Relation Networks [微软 2017]


ResNet
LR-Net[微软2019.4]

stage	output	ResNet-50	**LR-Net-50 (7×7, *m*=8)**
res1	112×112	7×7 conv, 64, stride 2	**1×1, 64** **7×7 LR, 64, stride 2**
res2	56×56	3×3 max pool, stride 2	3×3 max pool, stride 2
		1×1, 64 3×3 conv, 64 1×1, 256 ×3	1×1, 100 **7×7 LR, 100** 1×1, 256 ×3
res3	28×28	1×1, 128 3×3 conv, 128 1×1, 512 ×4	1×1, 200 **7×7 LR, 200** 1×1, 512 ×4
res4	14×14	1×1, 256 3×3 conv, 256 1×1, 1024 ×6	1×1, 400 **7×7 LR, 400** 1×1, 1024 ×6
res5	7×7	1×1, 512 3×3 conv, 512 1×1, 2048 ×3	1×1, 800 **7×7 LR, 800** 1×1, 2048 ×3
	1×1	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax
# params		**25.5**$\times10^{6}$	**23.3**$\times10^{6}$
FLOPs		**4.3**$\times10^{9}$	**4.3**$\times10^{9}$

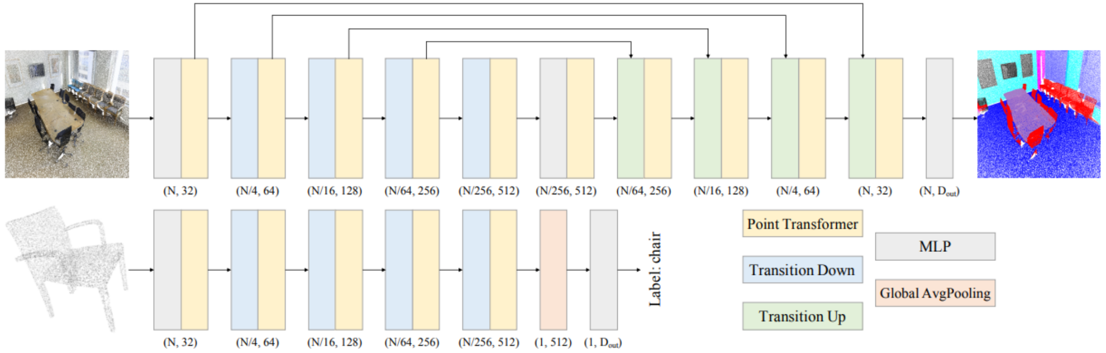
各种视觉问题取得性能突破

- 应用到视频领域
  - ViViT[谷歌]/MViT[Meta]/Video-Swin[微软]



ViViT

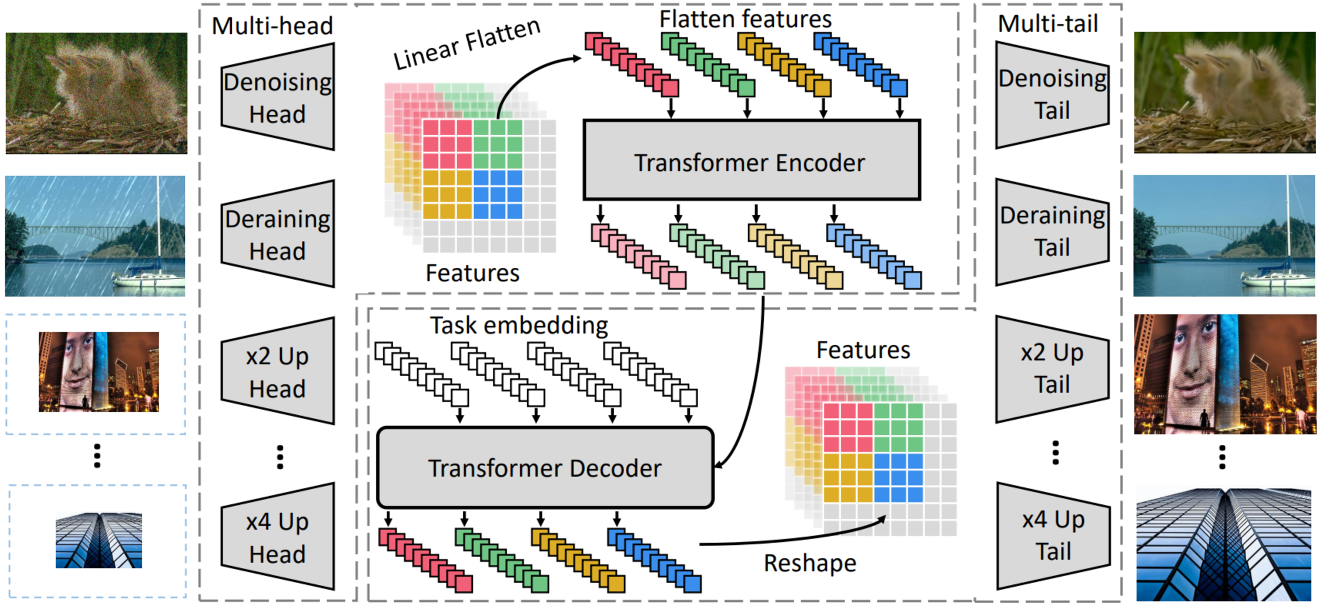
- 应用到3D视觉
  - Point Transformer [港中文&牛津]
  - Stratified Transformer [港中文&旷世]/SST[自动化所&图森]

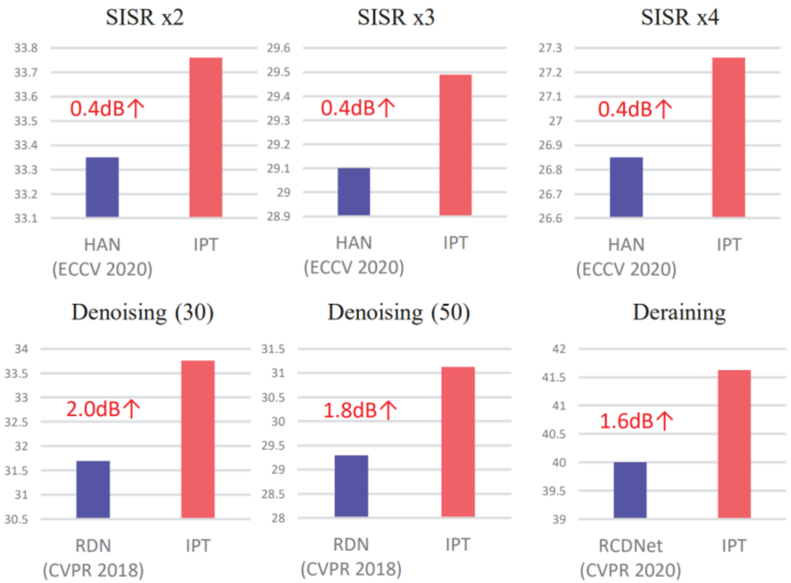


Point Transformer

各种视觉问题取得性能突破

- 应用到底层视觉
  - Image Processing Transformer[北大&华为]

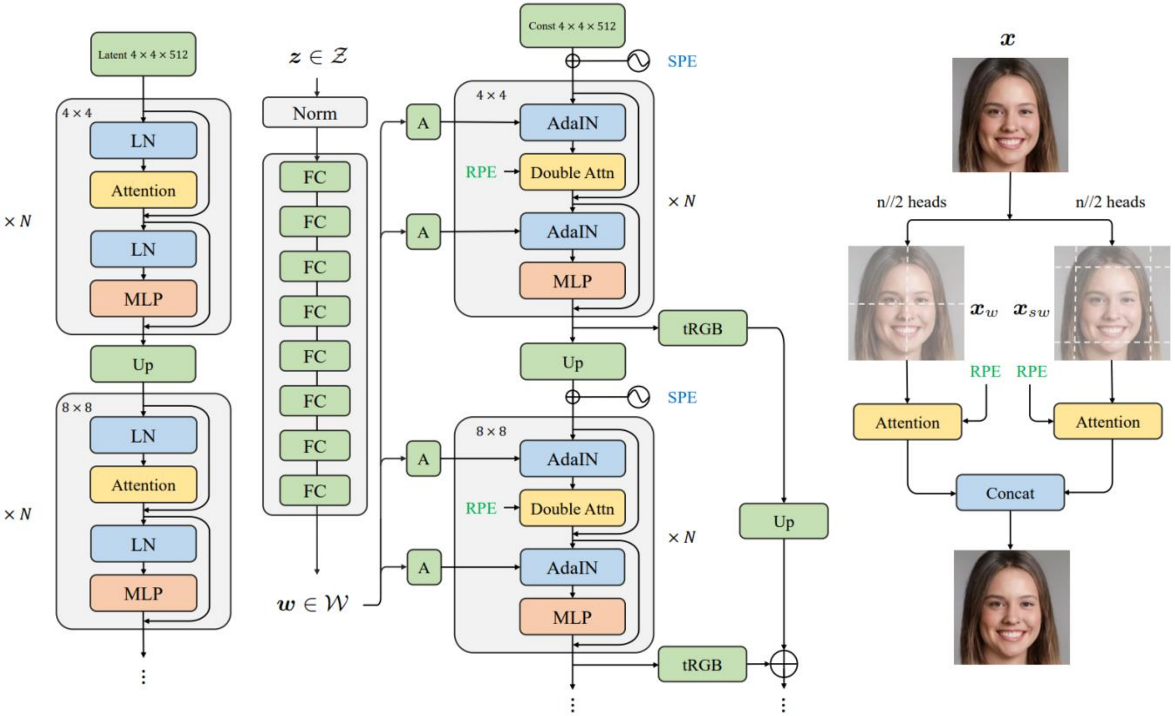




- 底层视觉其它代表性工作：
  - MAXIM[谷歌]/Uformer[中科大]/Restormer[UCM]/SwinIR[ETH]

各种视觉问题取得性能突破

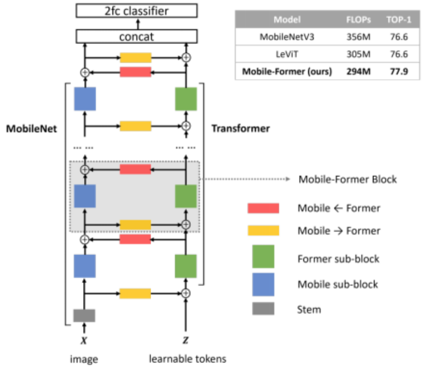
- 应用到图像生成领域
  - 首次尝试
    - TransGAN [UT Austin]
  - 首次超越基于CNN的GAN方法
    - StyleSwin[微软]



StyleSwin

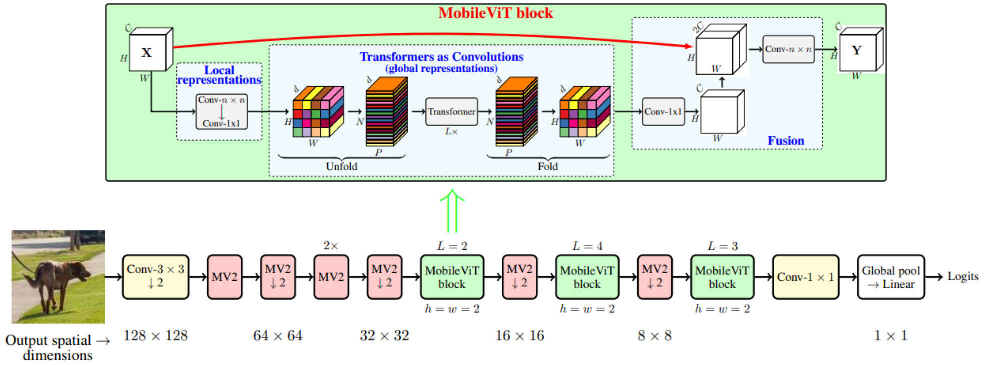
Transformer的轻量化与软硬件生态

- 轻量模型：
  - Mobile-Former[微软] / MobileViT [苹果]
  - MiniViT&TinyViT [微软]/NextViT [字节]/LVT [Adobe]



Model	FLOPs	TOP-1
MobileNetV3	356M	76.6
LeViT	305M	76.6
Mobile-Former (ours)	**294M**	**77.9**

Mobile-Former



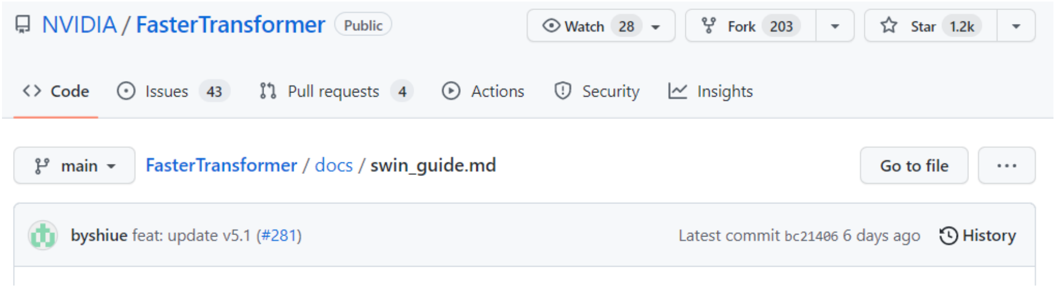
MobileViT

- 软硬件生态优化
  - Nvidia H100 / Nvidia-Swin

NVIDIA H100
TENSOR CORE GPU
Unprecedented performance, scalability, and
security for every data center.

H100 Transformer Engine Supercharges AI Training, Delivering Up to 6x Higher Performance Without Losing Accuracy

March 22, 2022 by DAVE SALVATOR



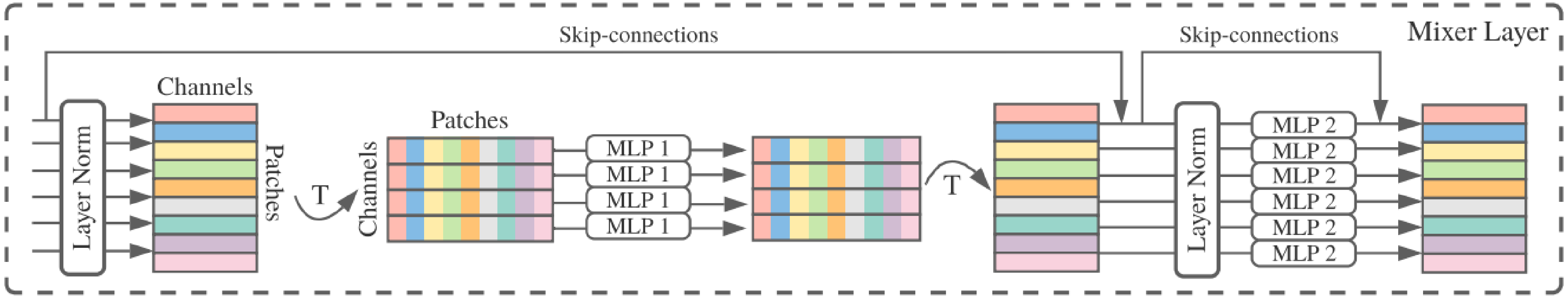
FasterTransformer / docs / swin_guide.md

byshiue feat: update v5.1 (#281)

Latest commit bc21486 6 days ago

启发更新型神经网络及传统卷积神经网络

- 受自注意力启发的高效空间特征融合机制
  - MLP-Mixer [谷歌]
  - 其它相关或者Follow-up工作：RepMLP[清华&旷视] /Swin-MLP [微软] /CycleMLP[港大]/ PoolFormer[SEA&NUS] /AS-MLP[上科大] / Hire-MLP[华为]

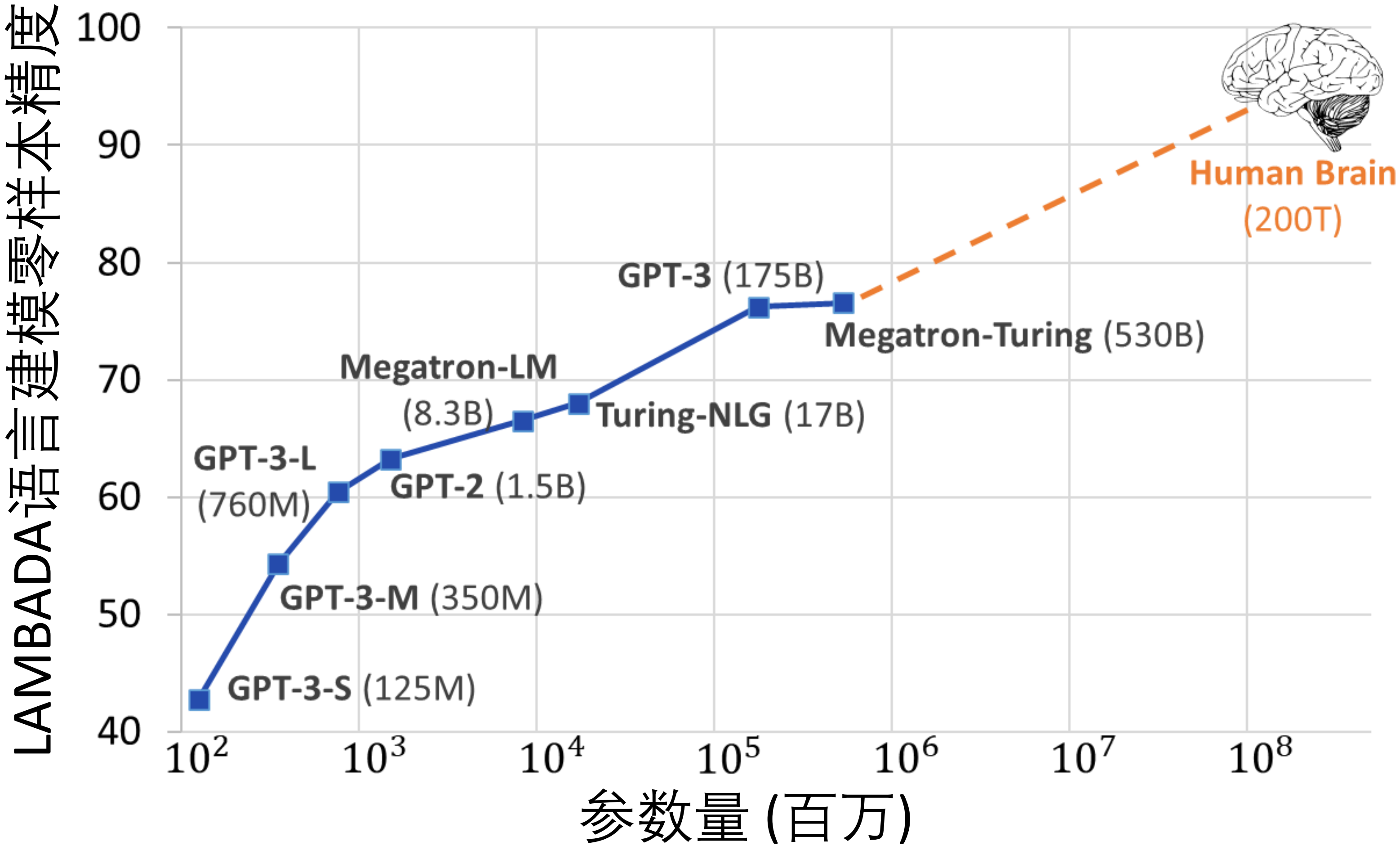


MLP-Mixer

- 卷积神经网络迎来“第二次增长”
  - D-DW Conv[南开&北大&微软]/ ConvNeXt[Meta] / RepLKNet[清华&旷视]

视觉Transformer大模型：视觉大模型能持续提升视觉任务的性能吗?

- **出发点**：NLP模型容量的扩展能持续改进NLP任务的性能



视觉Transformer大模型：视觉大模型能持续提升视觉任务的性能吗?

- Scaling ViT [谷歌]
  - 18亿参数（30亿标注图像）取得ImageNet-1K分类任务新纪录90.45%
  - 视觉Transformer同样符合Scaling Law（扩展定律）

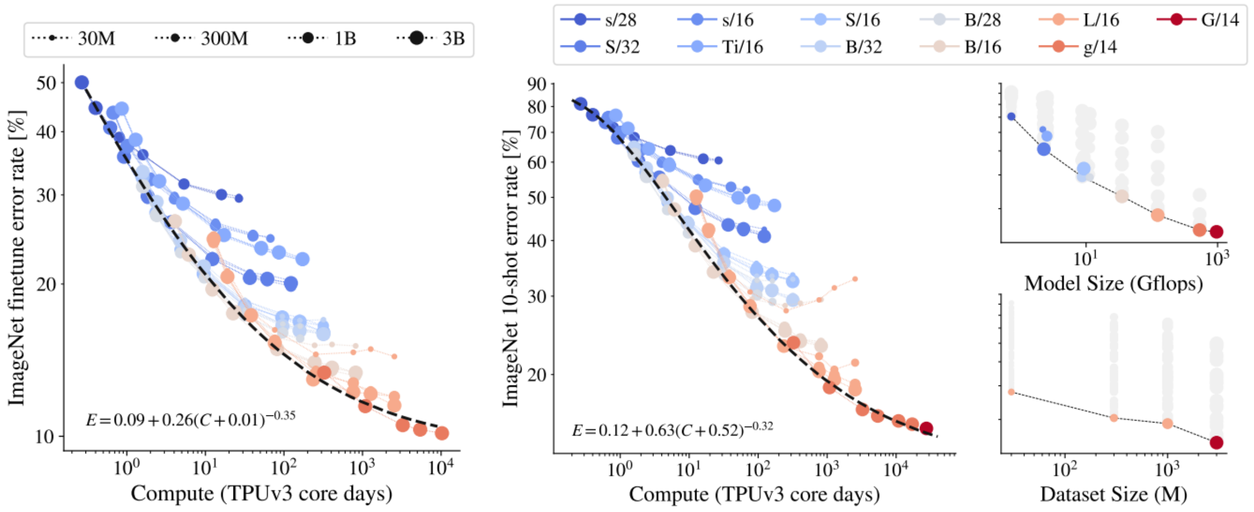


Figure 1: **Left/Center**: Representation quality, measured as ImageNet finetune and linear 10-shot error rate, as a function of total training compute. A saturating power-law approximates the Pareto frontier fairly accurately. Note that smaller models (blue shading), or models trained on fewer images (smaller markers), saturate and fall off the frontier when trained for longer. **Top right**: Representation quality when bottlenecked by model size. For each model size, a large dataset and amount of compute is used, so model capacity is the main bottleneck. Faintly-shaded markers depict sub-optimal runs of each model. **Bottom Right**: Representation quality by datasets size. For each dataset size, the model with an optimal size and amount of compute is highlighted, so dataset size is the main bottleneck.

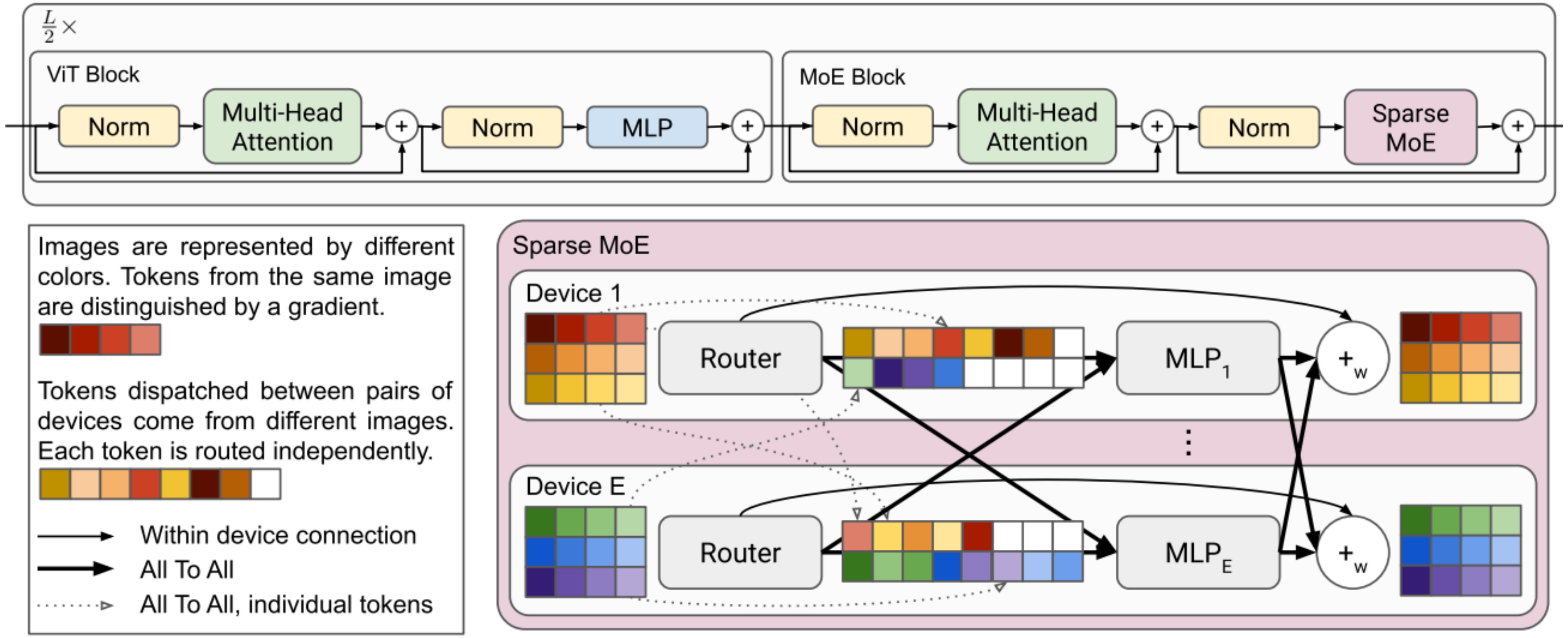
视觉Transformer大模型：视觉大模型能持续提升视觉任务的性能吗?

- Swin V2 [微软 2021.11]
  - 首次证明视觉大模型能显著帮助下游检测分割等任务，在四个任务上取得新的记录
  - 解决视觉大模型预训练和应用中的三大问题
    - 预训练稳定性问题
    - 解决预训练和下游分辨率差异的问题
    - 解决“数据饥饿”问题（比谷歌需要的标注数据小40倍）

模型	IN val v2 (图像分类)	COCO test-dev (物体检测)	ADE20K val (语义分割)	Kinetics-400 (视频分类)
Swin V1 (2021.3)	77.5	58.7	53.5	84.9
此前SOTA	83.3 (谷歌2021.7)	61.3 (MSRA, 2021.7)	58.4 (MSRA, 2021.10)	85.4 (谷歌, 2021.10)
Swin V2 (2021.11)	**84.0** **(+0.7)**	**63.1** **(+1.8)**	**59.9** **(+1.5)**	**86.8** **(+1.4)**

视觉Transformer稀疏大模型

- 出发点：人脑是稀疏模型，MoE是目前为止唯一能有效扩展到万亿参数的模型
- 视觉稀疏模型
  - 150亿参数的ViT-MoE模型 [谷歌]



$\frac{L}{2}\times$
ViT Block
Norm
Multi-Head Attention
+
Norm
MLP
+
MoE Block
Norm
Multi-Head Attention
+
Norm
Sparse MoE
+
Images are represented by different colors. Tokens from the same image are distinguished by a gradient.
Tokens dispatched between pairs of devices come from different images. Each token is routed independently.
Within device connection
All To All
All To All, individual tokens
Sparse MoE
Device 1
Router
MLP_1
$+_w$
⋮
Device E
Router
MLP_E
$+_w$

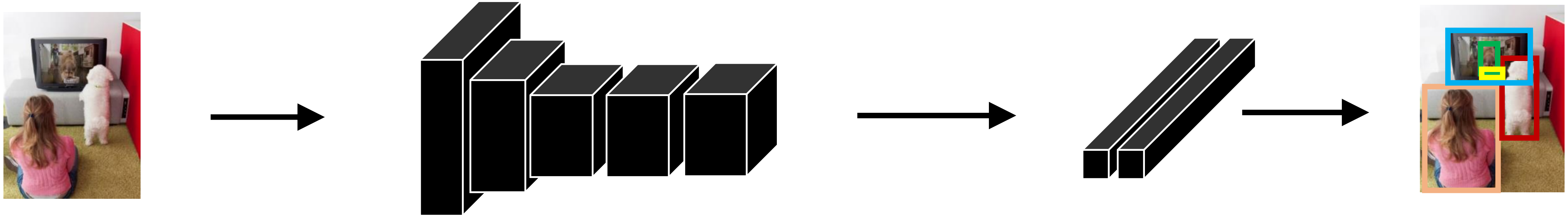
ViT-MoE

- 基于Tutel系统Swin-MoE [微软]
  - 代码和模型已开源

2021——视觉Transformer年： Transformer头部网络统一下游任务（变革#2）

视觉骨干网络 & 预训练

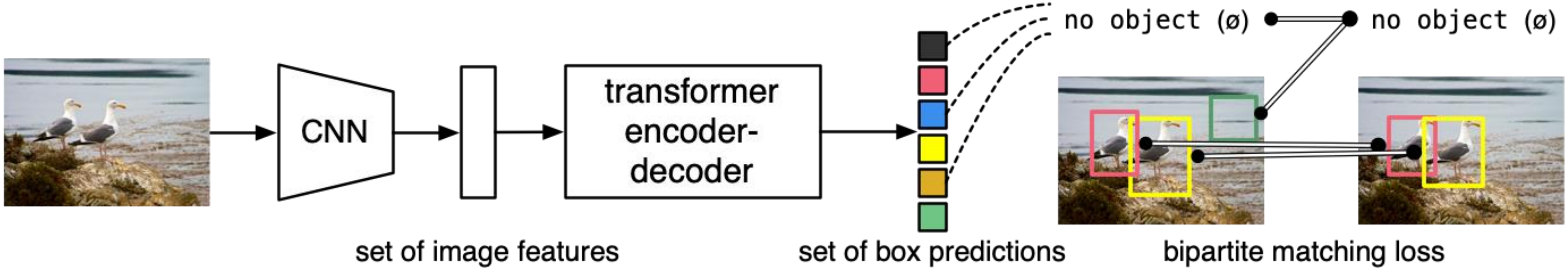
头部网络



CNN时代	CNN e.g., ResNet	有监督预训练 (图像分类)	针对不同任务的不同架构 e.g., Mask R-CNN, FCN
Transformer 时代	**Transformer** **e.g., ViT, Swin**	**自监督预训练MIM：** **e.g., BEiT/MAE/SimMIM**	**Transformer** **e.g., DETR**

针对下游任务的Transformer解码器网络

- 缘起：DETR [Meta 2020]
  - 物体检测：利用Transformer解码器将图像翻译成物体

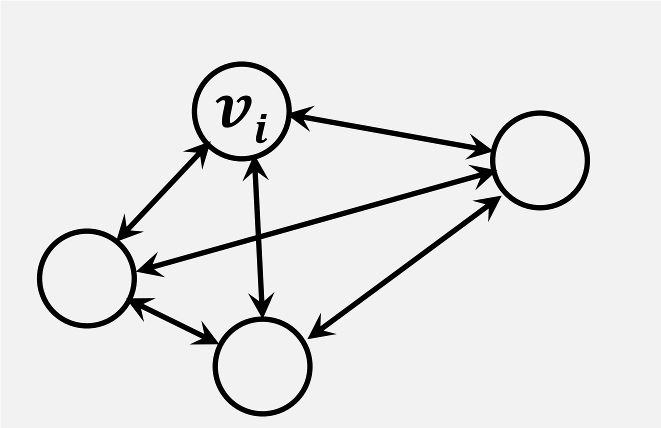


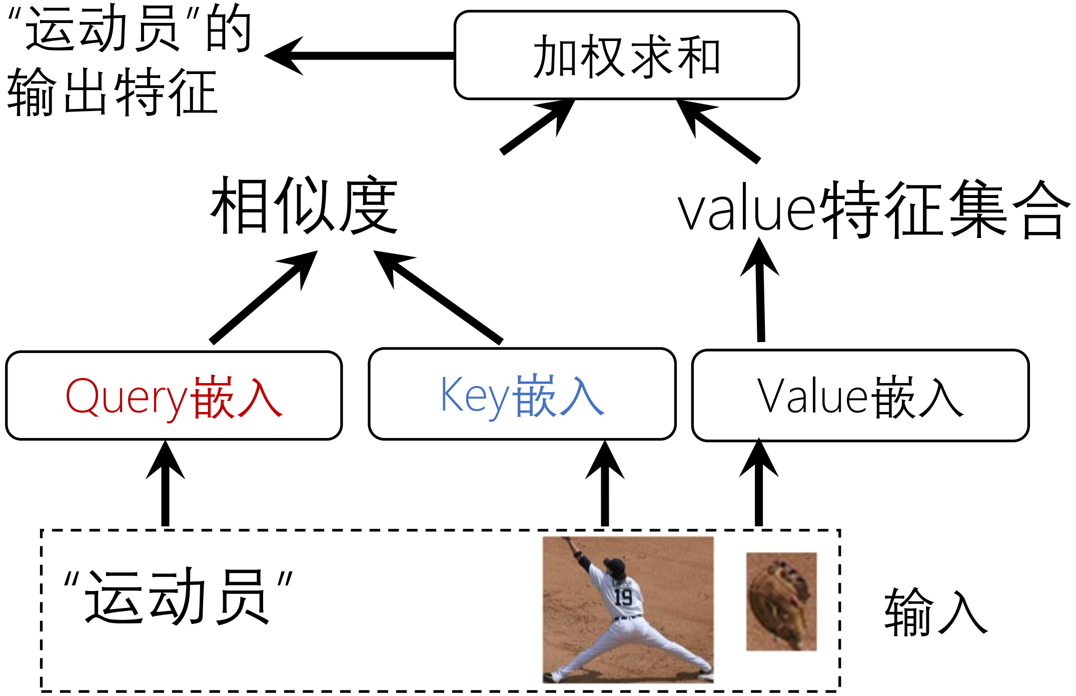
- 重要意义：**提供了一种统一各种下游感知任务的建模解法**
- 更早期工作
  - Relation Networks[微软 2017]：首个基于注意力的端到端物体检测器
  - Learning Region Features [北大&微软 2018]: 像素-区域&Transformer解码器

Nicolas Carion et al. End-to-End Object Detection with Transformers. ECCV2020

为什么Transformer解码器很通用?

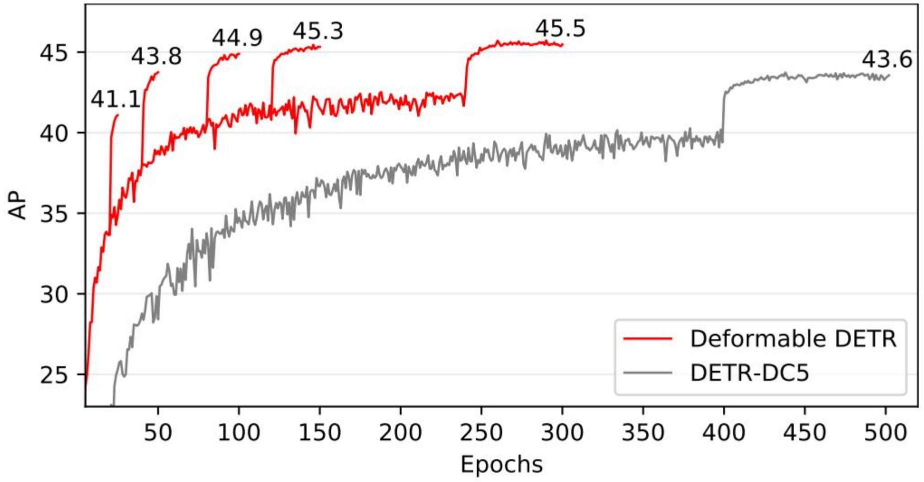
- 图的节点和边可以表示任意概念（实体或者抽象）以及他们之间的关系
- 验证的哲学去构建异构或同构实体间的关系

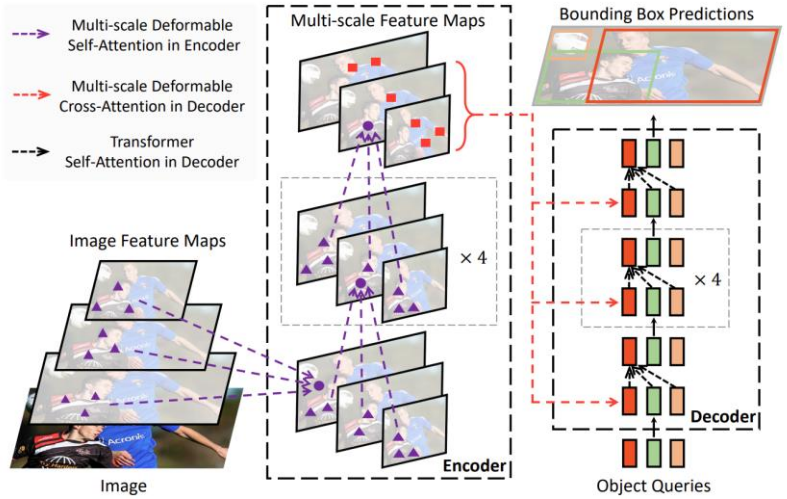




DETR的改进

- **加速 & 提升性能：**Deformable DETR [商汤 2021]





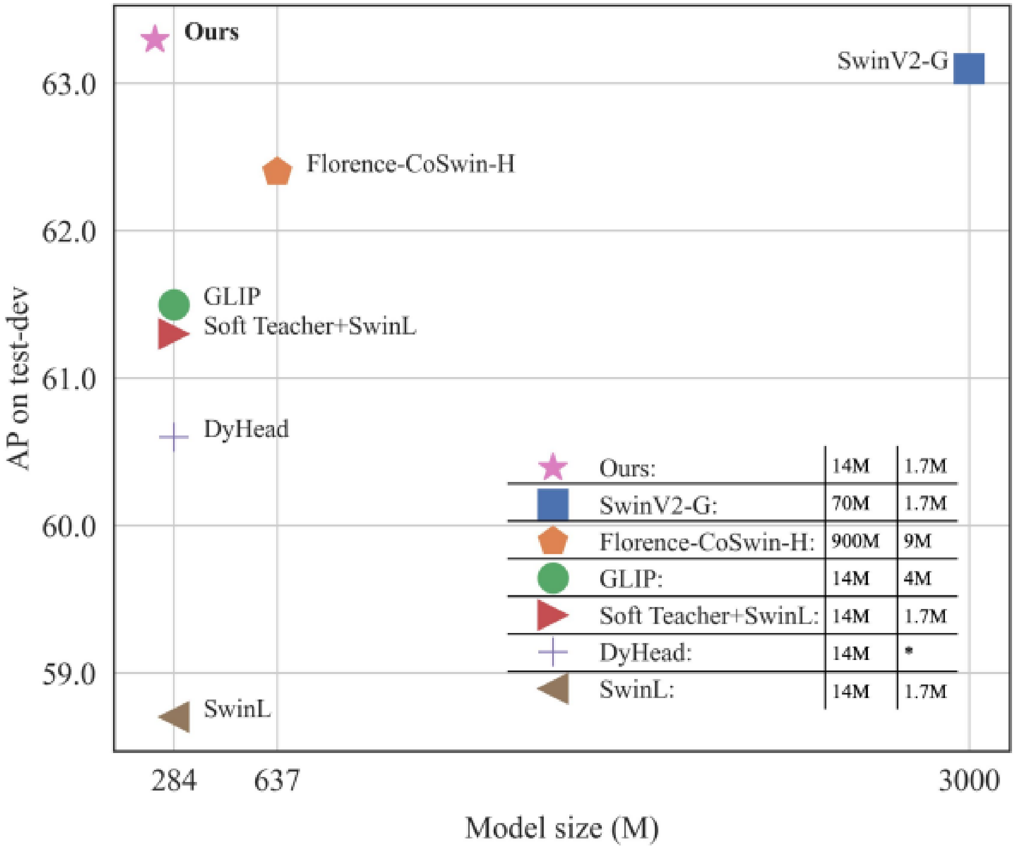
Multi-scale Deformable Self-Attention in Encoder
Multi-scale Deformable Cross-Attention in Decoder
Transformer Self-Attention in Decoder
Multi-scale Feature Maps
Bounding Box Predictions
Image Feature Maps
× 4
× 4
Encoder
Decoder
Image
Object Queries

- 其它相关工作
  - SMCA-DETR[上海AI实验室&CUHK&北大]/Conditional-DETR[中科大&北大&微软] / Conditional-DETRV2[北大&百度]/Anchor-DETR[旷视]/Efficient-DETR[旷视] / CF-DETR[北邮&华为]/FP-DETR[中科大&京东]/DAB-DETR[清华&IDEA] /DN-DETR[清华&IDEA]/DINO-DETR[清华&IDEA]/H-DETR[北大&中科大&微软]

Xizhou Zhu et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection. ICLR 2021

量变引起质变

- 取得性能里程碑：DINO[港科大&清华&IDEA 2022.3]
  - DETR框架首次在COCO检测上取得新纪录



★	Ours:	14M	1.7M
■	SwinV2-G:	70M	1.7M
⬟	Florence-CoSwin-H:	900M	9M
●	GLIP:	14M	4M
▶	Soft Teacher+SwinL:	14M	1.7M
+	DyHead:	14M	*
◀	SwinL:	14M	1.7M

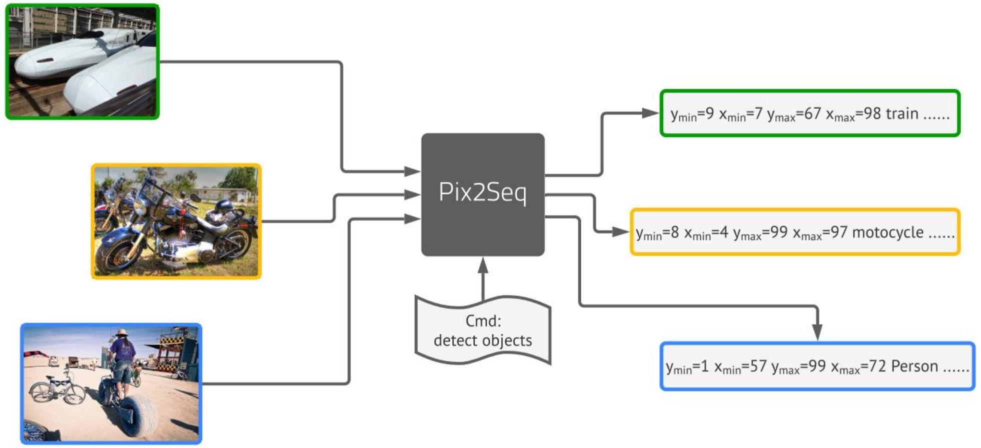
Hao Zhang et al. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arxiv 2022.

推广到更广泛的视觉问题

- 推广到分割问题
  - 图像分割 MaskFormer[UIUC&Meta]/Mask2Former[UIUC&Meta]
  - 视频分割 VisTR[美团]/IFC[延世大学&Adobe]/SeqFormer[华科&ByteDance]/VITA[延世大学&Adobe]
- 推广到3D检测领域
  - 室内场景：3DETR[Meta]/Group-Free-3D[MSRA]
  - 基于多视角图像：DETR3D[MIT&CMU&清华]/BEVFormer[南大&上海AI实验室&HKU]
PETR[旷视]/PETRv2[旷视]
- 推广到多模态
  - MDETR[NYU&FAIR]/FUTR3D[复旦&CMU&MIT&清华]
  - ViP3D[清华&CMU&MIT]/TransFusion[HKUST&华为&CUHK]

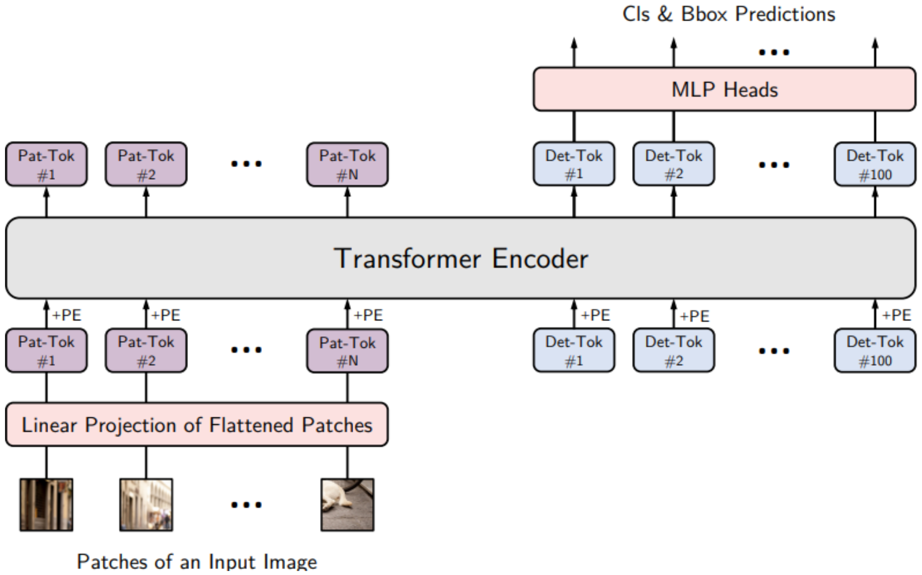
受DETR启发的新结构

- Pixel2SeqV1&V2[谷歌 2021]
- YOLOS [华科 2021]



Pix2Seq
Cmd: detect objects
y_{min}=9 x_{min}=7 y_{max}=67 x_{max}=98 train ......
y_{min}=8 x_{min}=4 y_{max}=99 x_{max}=97 motocycle ......
y_{min}=1 x_{min}=57 y_{max}=99 x_{max}=72 Person ......

Pixel2SeqV1

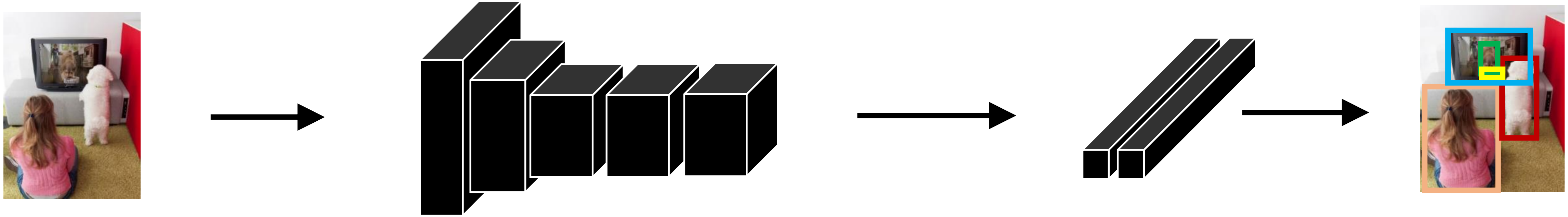


YOLOS

2021——视觉Transformer年： 从有监督预训练到自监督MIM（变革#3）

视觉骨干网络 & 预训练

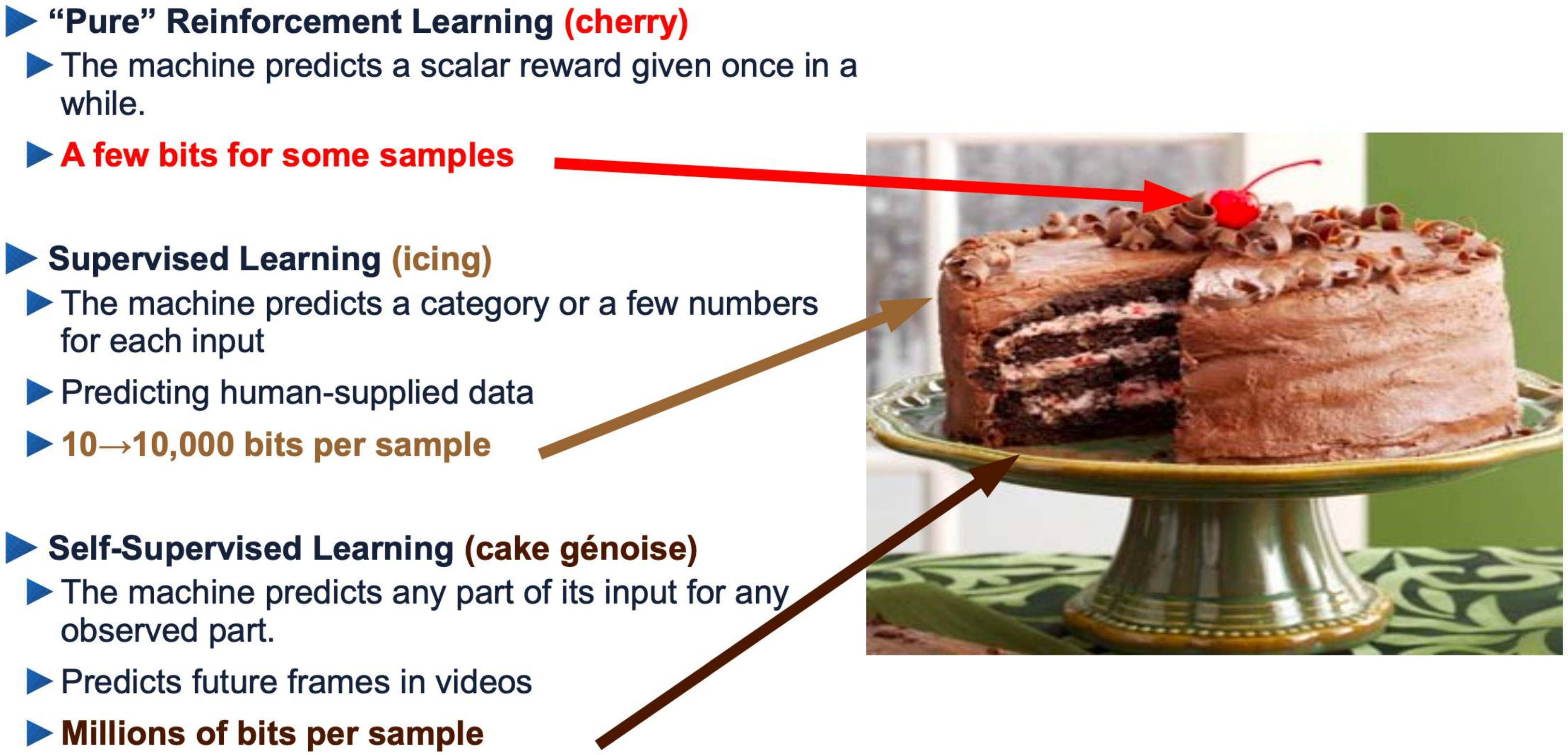
头部网络



CNN时代	CNN e.g., ResNet	有监督预训练 (图像分类)	针对不同任务的不同架构 e.g., Mask R-CNN, FCN
Transformer时代	Transformer e.g., ViT, Swin	**自监督预训练MIM： e.g., BEiT/MAE/SimMIM**	Transformer e.g., DETR

为什么需要自监督预训练？信息量的视角

- Yann LeCun的蛋糕类比：自监督学习能挖掘图像中的更多信息

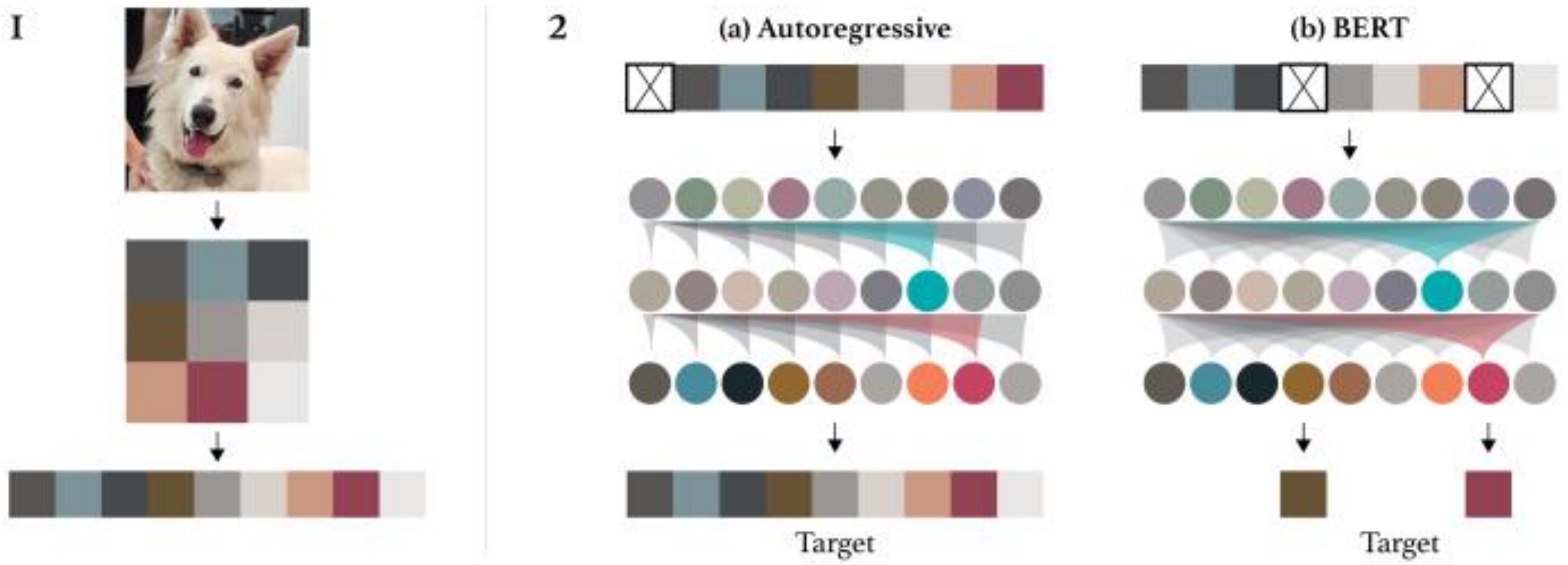


- **“Pure” Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

- Transformer的强大建模能力需要更多信息来“**填充**”

视觉Transformer的自监督预训练方法：掩码图像建模MIM

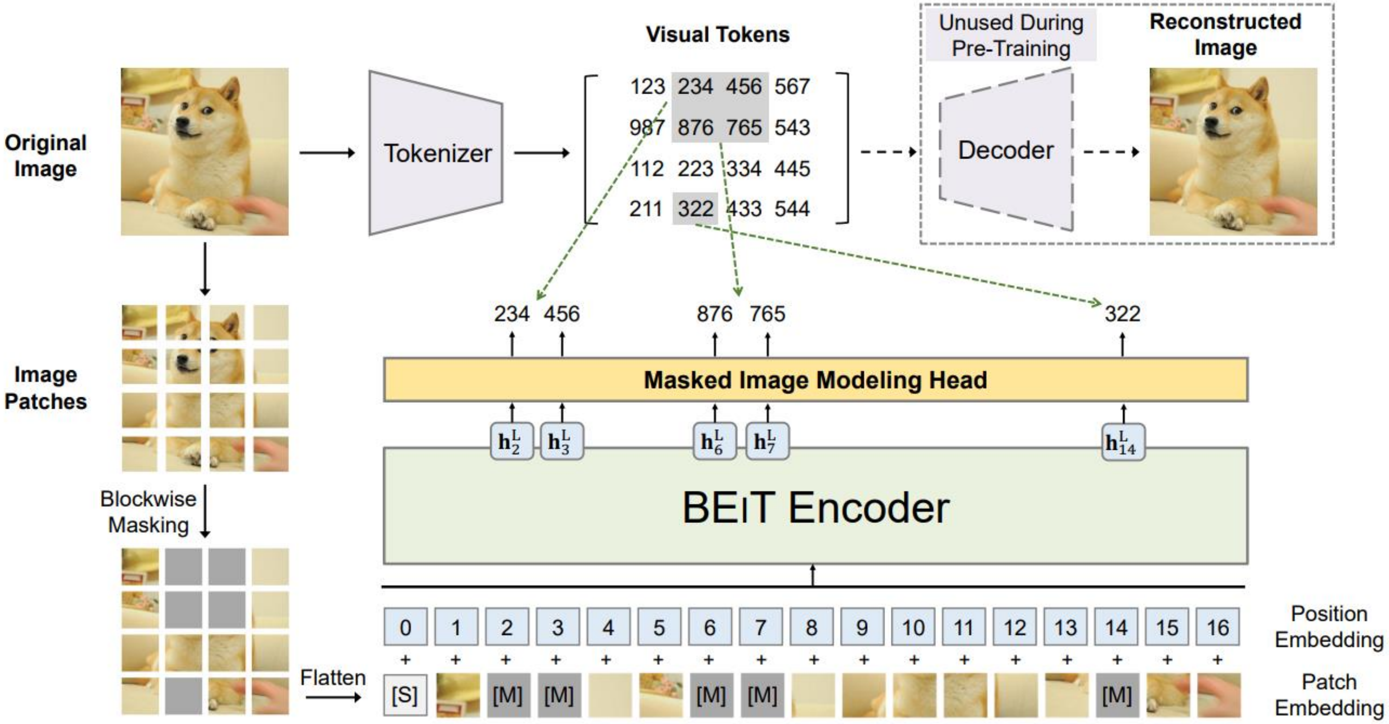
- 类BERT/GPT方法的先驱尝试
  - Image GPT[OpenAI 2020.6]



Mark Chen et al. Generative Pretraining from Pixels. ICML 2020.

掩码图像建模（MIM）性能里程碑

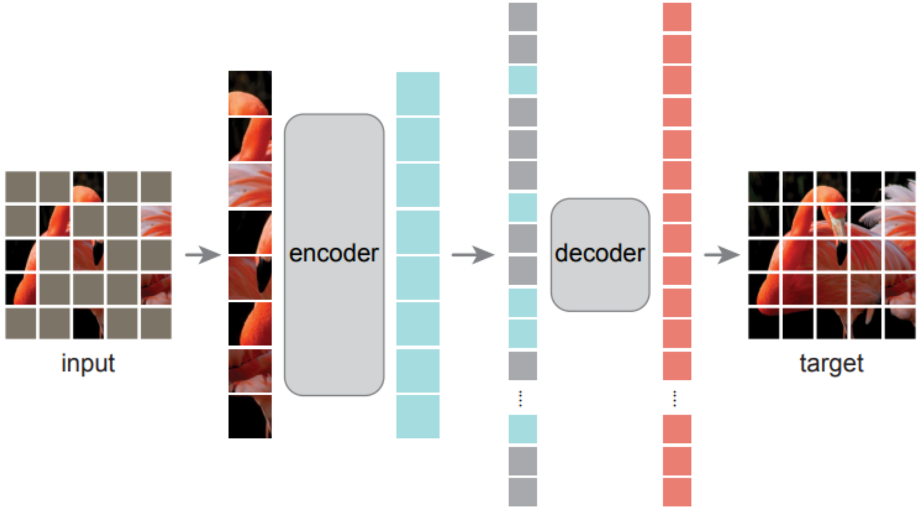
- 微调性能的突破：BEiT [微软 2021.6]
  - VQ-VAE得到视觉token，从而可以应用与BERT一模一样的方法



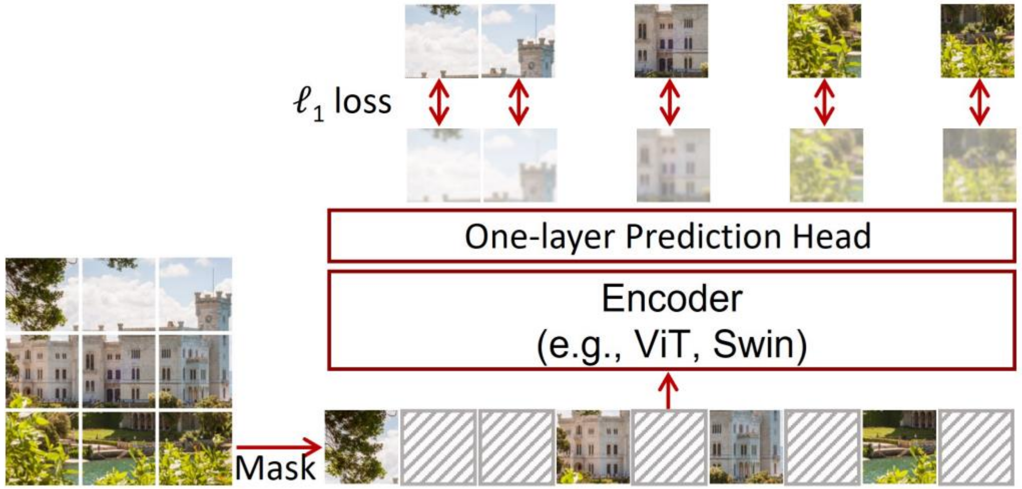
Hangbo Bao et al. BEiT: BERT Pre-Training of Image Transformers. ICLR 2022.

掩码图像建模（MIM）兴起高潮

- 兴起高潮（2021.10-11）： iBOT [字节2021.11] / MAE[Meta 2021.11] / SimMIM [微软2021.11]



MAE



SimMIM

- 百花齐放（2021年12月-)
  - 图像：data2vec [Meta] / PeCo[微软] / CAE[北大&百度]
  - 视频：MaskFeat [Meta] / BEVT [复旦&微软] / VideoMAE [南大/Meta]
  - 下游任务：ViTDet [Meta] / MIMDet [华科]

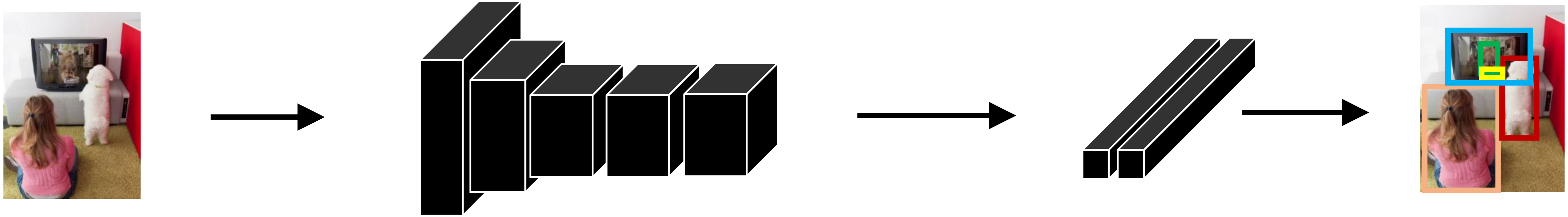
掩码图像建模（MIM）还在快速进展中...

- 继续提升MIM性能
- 理解MIM效果好的原理
- 在数据扩展上的表现
- 启发其它表征学习方法的改进（例如对比学习）

小结：2021——围绕视觉Transformer的三大变革：视觉Transformer年

视觉骨干网络 & 预训练

头部网络



CNN时代	CNN e.g., ResNet	有监督预训练 (图像分类)	针对不同任务的不同架构 e.g., Mask R-CNN, FCN
⇩	⇩	⇩	⇩
Transformer时代	**Transformer** **e.g., ViT/Swin**	**自监督预训练MIM：** **e.g., BEiT/MAE/SimMIM**	**Transformer** **e.g., DETR**

谢谢大家！